

## TRIFACTA 调查报告



Trifacta has brought an entirely new level of productivity to the way our analyst and IT teams work together to explore diverse data and define analytic requirements.



# 介绍

## 数据清洗：

现如今数据量正在急速增长，谁能够分析这些数据，得到有价值的信息，就能够成功。然而因为数据的复杂性，往往如果要分析数据，那么分析人员需要花费大概百分之八十的时间在数据整理上。一个成功的分析需要准确性，这就需要针对每一个任务有规范的结构化的数据。数据清洗就是把原始数据转化成格式化数据的一个过程。

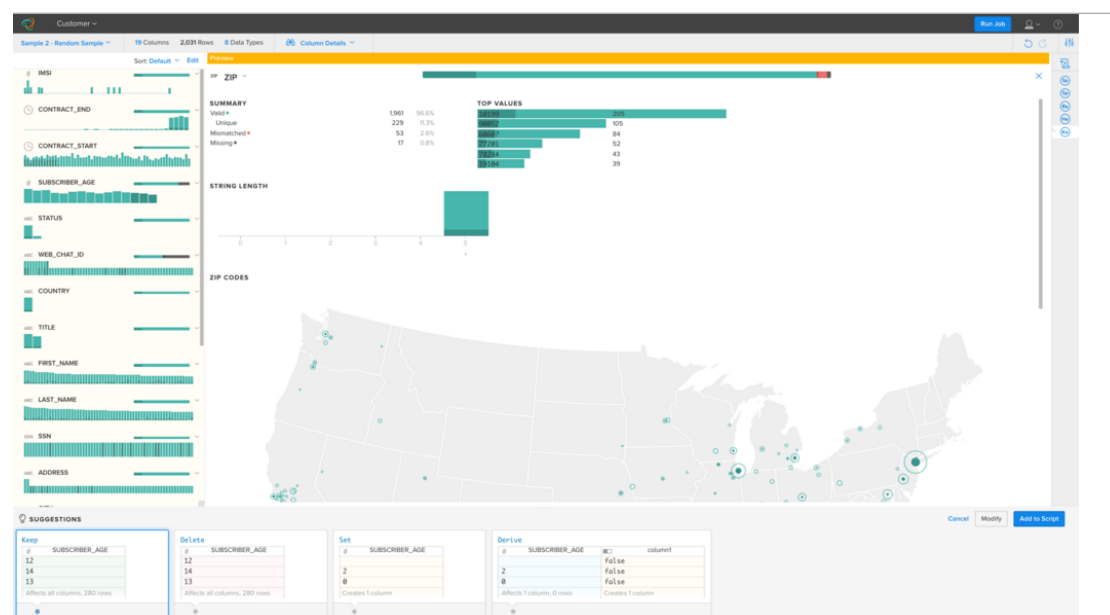
## Trifacta

Trifacta 能够让个人或者企业实现数据清洗的过程，不论你的目标是什么，你都需要数据清洗的过程，而 Trifacta 提供了一个数据清洗的平台，让用户能够更快，更直观，更有效的收集、清理和转换数据。而且用户不必是程序工程师，Trifacta 的界面完全是交互式的，而不是代码式的。这使得分析师和数据专家将能使用可视化的方式去清洗数据集。

# Trifacta 提供的功能

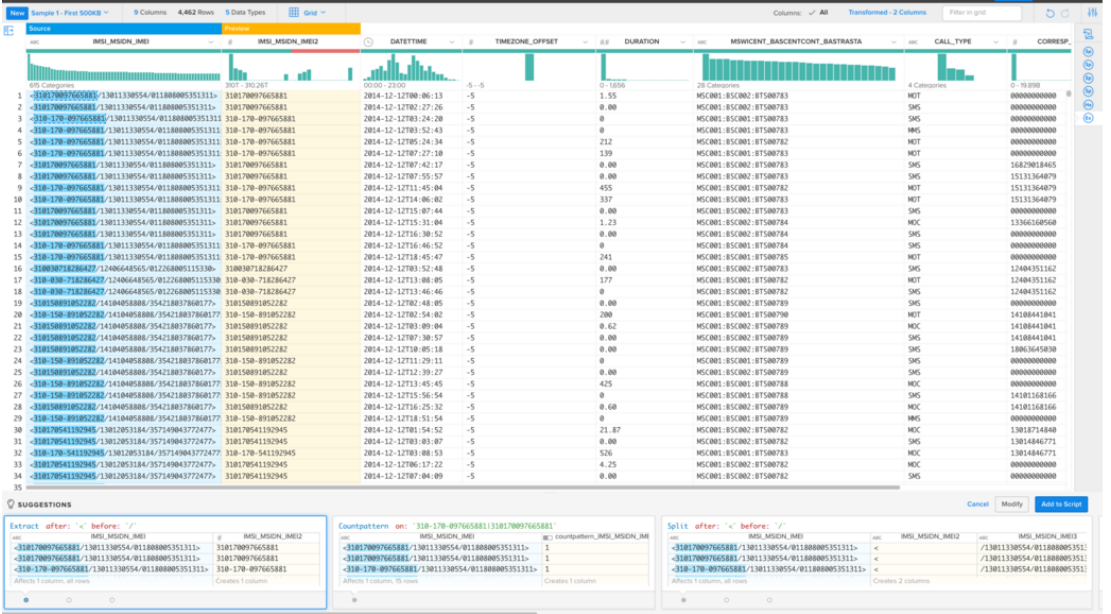
## 发现

能够找到原始数据的使用价值和有用的信息关键是弄清楚你的数据到底是什么，数据长什么样子，和数据可能有哪些用途。这一功能能够让你更好地了解一些数据的特征，例如数据的值的分布，和一些对于数据处理转换和分析的建议。



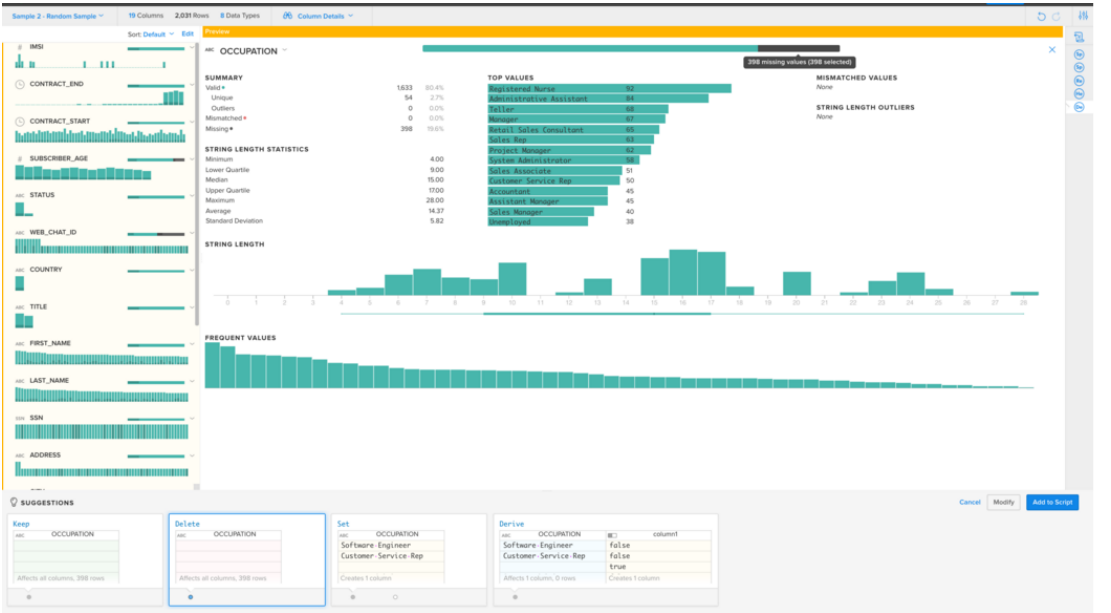
结构化

数据结构化，这一功能是必须的。原始数据来自不同的领域，可能有不同的编码，格式，数据类型，导致用户根本无法阅读，就别提什么分析了。所以本软件提供了结构化数据的功能。



清理

数据清理能够去掉数据中的错误数据和丢失数据，整理对于统一数据不同的表达，例如 CA, Cal 和 Calif。



## 丰富

该功能能够让用户选择不同的数据集，进行数据融合。

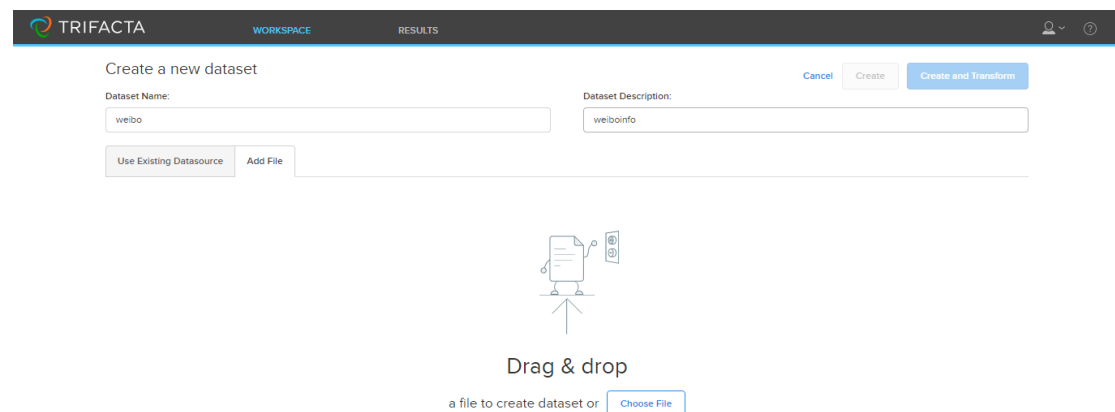
## 统计和导出

能够对清洗的结果进行统计展示，用于验证。然后能对结果进行导出

以上几种功能，Trifacta 都应用机器学习算法为重新组织信息和整理提供建议。分析师可以将数据集分组为信息的逻辑部分，每次将其规范化，并在其工作过程中以友好的界面方式显示。Trifacta 的目标客户不仅包括普通的商业用户，数据科学家也是其争取的对象。用户可用 Trifacta 利用多种可视化数据方式来浏览这些数据。应用还可以对你下一步的操作提出若干建议，操作执行的效果还可以预览。一旦决定了希望执行的操作，相应的代码或查询就可以被生成执行。

## 使用情况

## 创建收集数据集



Trifacta WORKSPACE RESULTS

Create a new dataset

Dataset Name: weibo

Dataset Description: weiboinfo

Cancel Create Create and Transform

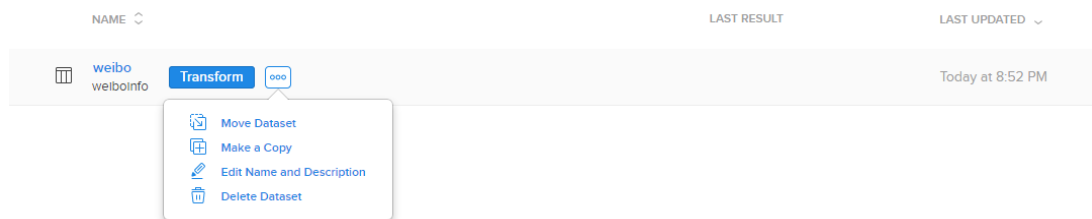
Use Existing Datasource Add File

Drag & drop  
a file to create dataset or Choose File

软件可以创建一个数据集，可以定义数据集的名称和描述信息。创建的数据有两种。一种是选择已有的数据源，一种是选择文件。选择后，将会显示选择了文件的信息，包括文件名称大小和创建时间。

Name	Size	Created
weibo.txt	61B	Today at 8:52 PM

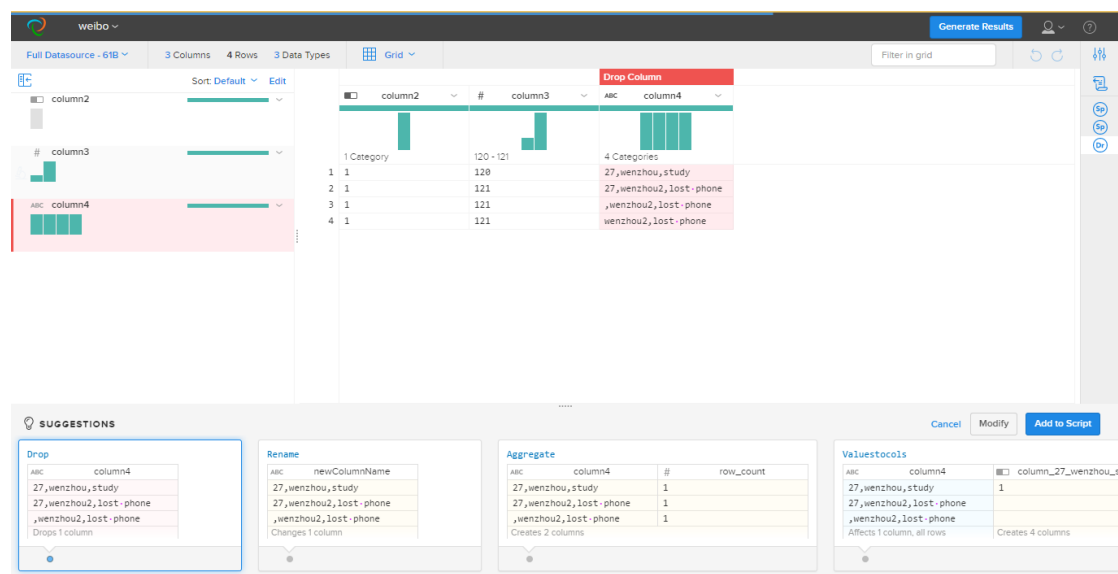
对于已经收集的数据集，软件提供的操作有备份，更改数据集名称和描述，删除等操作



## 数据集的清洗分析和转化

单击数据集旁边的按钮可以对数据集整理。

整理后的界面长成下面的样子：



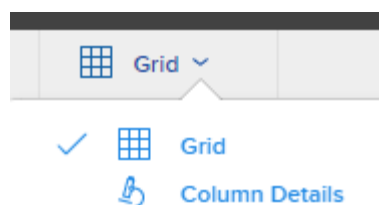
## 整体统计

工具栏有两条：

横向的工具栏：



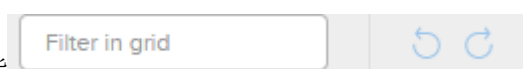
展示了数据集的大小，行数，列数，数据类型种类数，然后提供了数据列查看的两种方法，

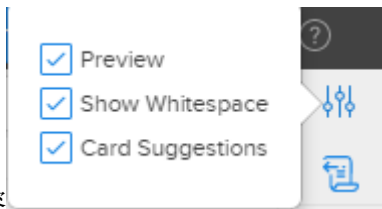
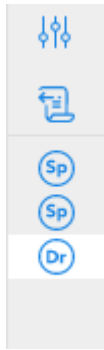


表格展示和详细展示。

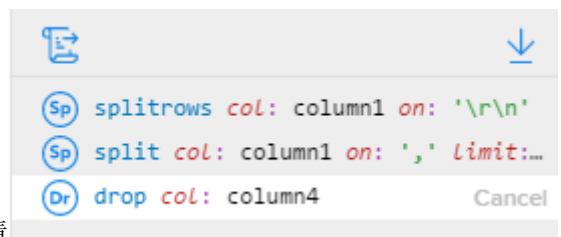
然后提供了数据过滤功能和 undo，redo 功能

纵向的工作栏：





主要的功能第一个图标是选择表格展示的内容

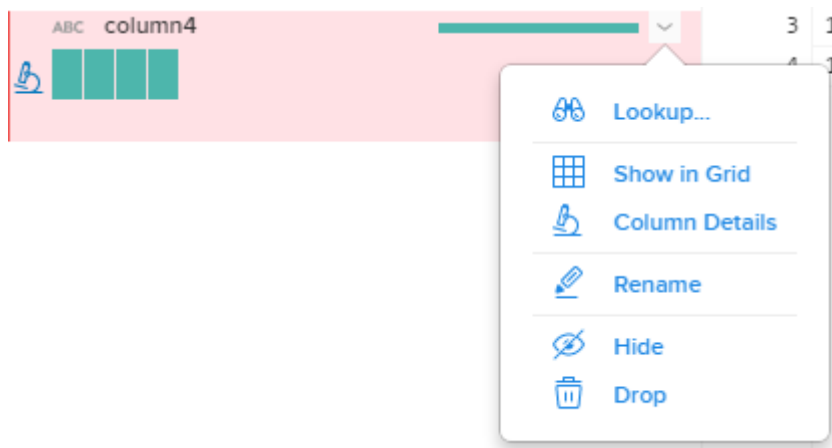


第二个图标展示用户每一步的操作日志详情

下面的图标是用户每一步的历史操作，可做回滚。就是历史列表。

## 列分析界面

左上的界面是对每一列的分析，我们以最后一列举例，对于该列可以进行的操作有查看原始数据，查看列的详细信息或者在表格中查看，重命名，隐藏删除等：

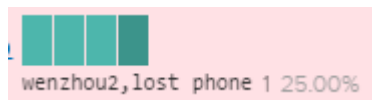


- ABC 编码了数据类型：

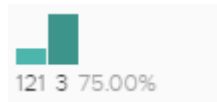


- 下面四个竖长条编码了数据的分布，而且当鼠标移动上去的时候会显示该长条的数值、

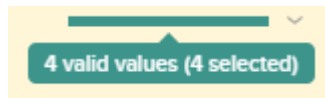
频数和频率：



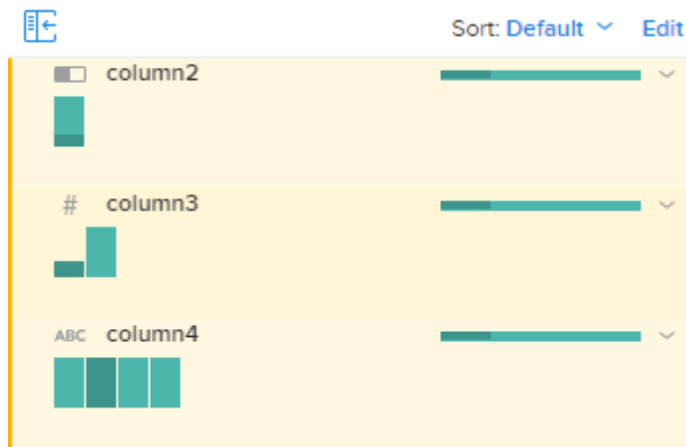
长度编码相对大小例如（深绿色为高亮）：



- 横长条编码了数据“干净”程度（有多少有效等）



- 在列列表中的高亮，点击后在所有列中会高亮相对应的地方。



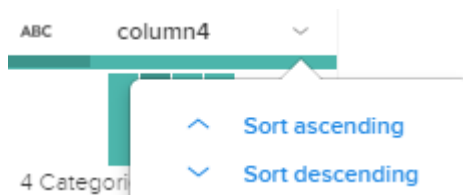
## 数据界面

数据界面和列界面的高亮是联动的。

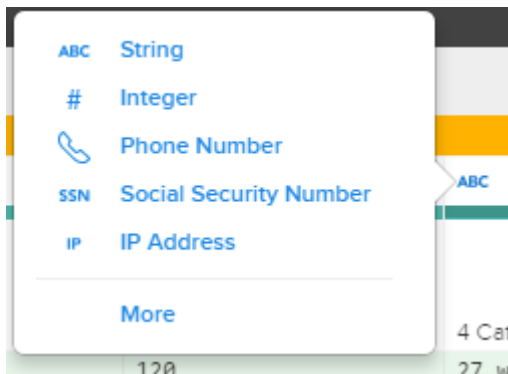
表格视图：

Preview			
	column2	# column3	ABC column4
	1 Category	120 - 121	4 Categories
1	1	120	27,wenzhou,study
2	1	121	27,wenzhou2,lost phone
3	1	121	,wenzhou2,lost phone
4	1	121	wenzhou2,lost phone

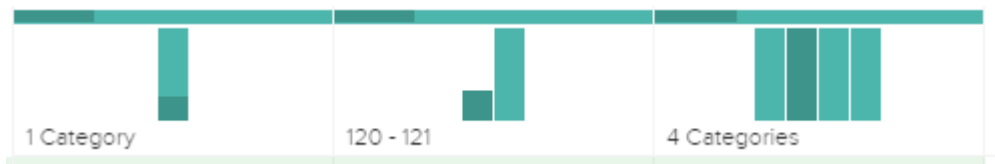
- 可以提供根据该列进行升序或者降序排序



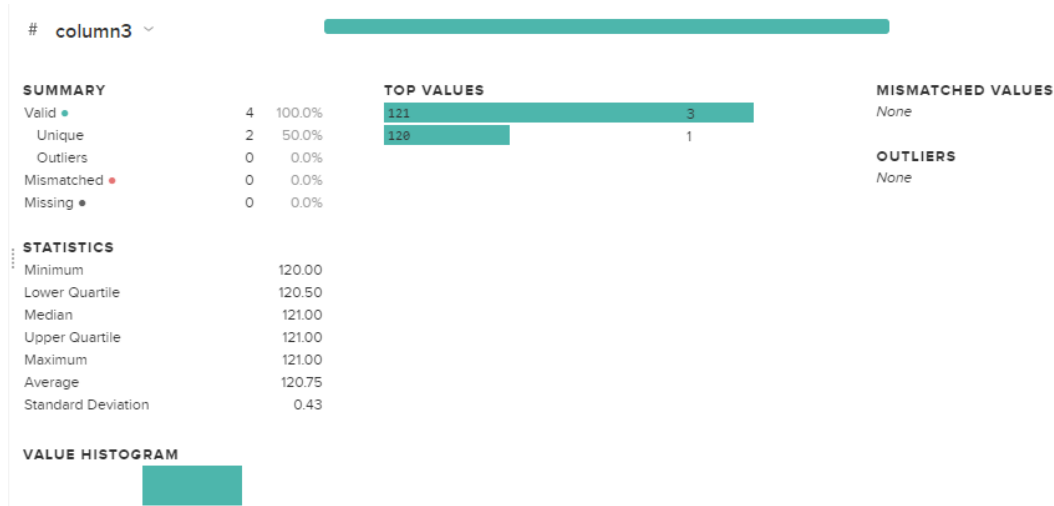
- 每一列商业显示了数据类型。数据类型可以进行更改：



- 同样展示了数据统计信息，其中多了一个数值范围：



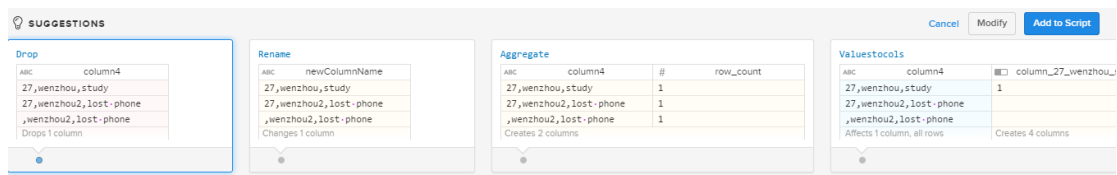
详细界面：



列举了该列的详细信息，尤其是一些详细的统计信息。

## 操作提示界面

对于列界面和数据界面中的数据或者列进行点击，可以进行一些数据清理的操作。用户不仅可以对整列或者对单个单元格进行修改，还可以点击统计信息来对符合条件的数据进行操作。操作界面列举了系统提供的很多操作，包括删除，重命名以及聚合，数数，设为 key 等的数据库操作。



用户可以根据图形化的界面点击选择自己所需要的操作。也可以点击 **modify** 按钮直接修改代码





进行操作后，历史列表将会记录，并提供 redo，undo 操作。  
用户还可以根据历史列表在任意两步操作之间加入一步操作。

## 结果导出

软件提供了对清洗后的数据导出的功能，可以导出不同格式的数据并，可以以后在软件的结果栏中对结果进行查看：

